

Beyond mIoU: A Deployment-Oriented Evaluation Framework for UAV Semantic Segmentation

Nadeem Gul

The School of Optoelectronics Engineering, Xidian University, Xi'an 710071, China

Meng Qin

The School of Optoelectronics Engineering, Xidian University, Xi'an 710071, China

Aamir Naveed Abbasi

The School of Artificial Intelligence, Xidian University, Xi'an 710071, China

Sichun Zheng

The School of Optoelectronics Engineering, Xidian University, Xi'an 710071, China

Gu Yuanyuan

The School of Optoelectronics Engineering, Xidian University, Xi'an 710071, China

Changqing Cao

chqcao@mail.xidian.edu.cn

The School of Optoelectronics Engineering, Xidian University, Xi'an 710071, China

Research Article

Keywords: UAV semantic segmentation, evaluation framework, Segmentation Efficiency Index (SEI), deep learning, remote sensing

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10997782/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Beyond mIoU: A Deployment-Oriented Evaluation Framework for UAV Semantic Segmentation

(Corresponding author: Changqing Cao)

Nadeem Gul¹, Meng Qin¹, Aamir Naveed Abbasi², Sichun Zheng¹, Gu Yuanyuan¹ and Changqing Cao¹

1.The School of Optoelectronics Engineering, Xidian University, Xi'an 710071, China.

2.The School of Artificial Intelligence, Xidian University, Xi'an 710071, China.

(e-mail: nadeem.gul@stu.xidian.edu.cn, 24191214608@stu.xidian.edu.cn, aamirnaveed@stu.xidian.edu.cn, 24191214586@mail.xidian.edu.cn, 25191111486@stu.xidian.edu.cn, chqcao@mail.xidian.edu.cn).

Abstract

Despite significant advances in UAV semantic segmentation, region-overlap accuracy remains the primary focus of current evaluation protocols, with mean Intersection over Union (mIoU) serving as the predominant benchmark. However, such metrics provide only a partial assessment of real-world deployment, as they overlook computational efficiency, robustness to image degradations, prediction calibration, and the spatial characteristics of segmentation errors. To overcome these limitations, we propose a deployment-oriented evaluation framework that integrates efficiency, robustness, calibration, and spatial error analysis for comprehensive assessment and model selection in UAV semantic segmentation. Using the Semantic Drone Dataset, consisting of 400 high-resolution images categorized into seven land-cover classes, we systematically train and evaluate eight configurations of U-Net, U-Net++, DeepLabV3+, FPN, PSPNet, and MAE under an identical experimental protocol. The proposed framework introduces the Segmentation Efficiency Index (SEI), a deployment-weighted metric that combines normalized accuracy, inference speed, and parameter count. The SEI demonstrates that the model achieving the highest accuracy may not necessarily represent the optimal choice under practical deployment constraints. Furthermore, we evaluate model robustness against 10 realistic image corruptions with predefined severity levels to quantify performance degradation under challenging imaging conditions. Moran's I-based spatial error analysis reveals statistically significant clustering patterns in segmentation errors, providing insights beyond conventional accuracy metrics. Additional evaluations of ensemble diversity, calibration performance, boundary transitions, and prediction stability further characterize model reliability. Overall, this study provides a comprehensive and deployment-oriented evaluation framework that enables scenario-specific selection of reliable UAV semantic segmentation models for practical applications.

Keywords: UAV semantic segmentation; evaluation framework; Segmentation Efficiency Index (SEI); deep learning; remote sensing.

1. Introduction

The widespread use of unmanned aerial vehicles (UAVs) equipped with high-resolution cameras has transformed remote sensing in several fields. Compared to satellite systems, UAV imagery has greater spatial resolution, operational flexibility, acquisition cost and user-defined temporal revisit intervals (Lyu et al., 2020). These properties make UAVs especially suitable for applications that need extensive investigation of the ground, such as urban planning, environmental monitoring, precision agriculture, disaster response and infrastructure inspection.

Semantic segmentation, i.e. assigning a class name to each pixel, is a critical step to extract meaningful information from UAV imagery. Segmentation classifies pixels to land-cover types such as buildings, vegetation, roads and water bodies, allowing for automatic mapping, change detection and quantitative analysis of urban and natural landscapes. Deep learning has revolutionised this task: encoder-decoder networks (Ronneberger et al., 2015), atrous-convolution models (L.-C. Chen et al., 2018), pyramid architectures (Lin et al., 2017; Zhao et al., 2017) and more recently transformer-based architectures (Liu et al., 2021; Xie et al., 2021) achieve accurate results on standard benchmarks.

However, the dominant evaluation paradigm has a key flaw: it is overly dependent on mIoU as a primary frequently the only metric of quality. This trend is shown by recent systematic reviews of UAV segmentation algorithms and datasets (J. Cheng et al., 2024). Each of the following dimensions has been studied in isolation but rarely considered jointly in the UAV segmentation literature. Robustness benchmarks are available for natural and driving imagery (Hendrycks & Dietterich, 2019; Kamann & Rother, 2020; Michaelis et al., 2019). Accuracy speed trade-offs are central to the efficient-segmentation literature (Xu et al., 2023; Yu et al., 2018). Calibration has been studied for classification (Guo et al., 2017). Spatial error characterisation appears in remote sensing quality assessment. However, earlier work has not brought together these aspects into a broad evaluation process on drone images, leaving practitioners without principled advice on concerns such as: How do models degrade under fog, rain, noise, or lighting variations during flight? What is the accuracy-speed-complexity trade-off, and can one measure it? Are errors spatially random or clustered in patterns amenable to targeted post-processing? How stable are predictions under the small viewpoint perturbations inherent to drone motion? Which class boundaries are most error-prone, and do models exhibit directional biases there? Can architectural diversity be exploited through ensembling?

To close this gap, we propose a comprehensive multi-dimensional evaluation framework for semantic segmentation of UAV aerial imagery and apply it to eight state-of-the-art model configurations as case studies to validate the effectiveness of the proposed framework, including U-Net (ResNet-34/50), U-Net++ (ResNet-34), DeepLabV3+ (ResNet-50/101), FPN (ResNet-50), PSPNet (ResNet-50), and MA-net (ResNet-50), using the Semantic Drone Dataset with seven grouped land-cover classes. The principal contributions of this study are three-fold. (1) We propose the Segmentation Efficiency Index (SEI), a composite metric that unifies normalized

accuracy, inference speed, and model complexity under configurable priority scenarios, enabling deployment-aware model selection that mIoU alone cannot support. (2) We present a systematic corruption robustness analysis for drone segmentation covering ten precisely parameterized real-world degradation types, showing 20–27% performance loss under adverse conditions. (3) We introduce spatial error autocorrelation analysis using Moran's I supported by statistical significance testing together with an integrated suite of complementary analyses (prediction stability under simulated drone motion, class transition errors, calibration, uncertainty and agreement mapping, minimum detectable object size, and ensemble diversity) that collectively reveal insights invisible to conventional metrics.

The rest of this paper is organised as follows Section 2 discusses related work; Section 3 introduces the suggested methodology (containing the dataset, architectures, the SEI metric and the assessment system); Section 4 presents the experimental set-up; Section 5 presents the findings. Section 6 examines findings, recommendations and limits. Section 7 closes the study.

2. Related Work

2.1 Semantic Segmentation Architectures

Semantic segmentation has been rapidly developed since Fully Convolutional Networks (FCN) have shown that classification networks may be converted to dense pixel-wise prediction (Long et al., 2015). However, FCN generates coarse borders due to aggressive down sampling. U-Net proposed an encoder–decoder structure with skip connections to maintain fine spatial detail (Ronneberger et al., 2015). However, the plain skip pathways in U-Net can propagate semantically shallow features. U-Net++ reduced the semantic gap by adding nested and dense skip connections (Zhou et al., 2018), however this required more processing. DeepLabV3+ [12] use atrous spatial pyramid pooling (ASPP) and a light decoder for multi-scale context and boundary refinement, however its deep backbones are computationally expensive. Feature Pyramid Networks (FPN) are originally created for detection (Lin et al., 2017) and may effectively utilise top-down multi-scale feature fusion, however they are not specially designed for dense aerial scenes. PSPNet (Zhao et al., 2017) aggregates the global context via pyramid pooling, which can be detrimental to fine-grained localisation. MAnet uses multi-scale attention in the decoder to highlight the informative features (Fan et al., 2020), but the vast number of parameters limits the edge deployment. More recent work has shifted toward transformers and efficiency-oriented designs. SegFormer couples a hierarchical transformer encoder with a lightweight MLP decoder (Xie et al., 2021); Swin Transformer provides a general purpose hierarchical attention backbone (Liu et al., 2021); Mask2Former unifies semantic, instance, and panoptic segmentation through masked attention (B. Cheng et al., 2022); and promotable foundation models such as the Segment Anything Model open new zero-shot possibilities (Kirillov et al., 2023), recently extended to video (Ravi et al., 2024) and adapted to remote sensing through prompt learning (Chen et al., 2024). State-space (Mamba) architectures have also entered remote sensing segmentation, offering linear-complexity global

context(Ma et al., 2024; Zhu et al., 2024). In parallel, real-time networks such as BiSeNet (Yu et al., 2018) and PIDNet (Xu et al., 2023) pursue explicit accuracy–latency trade-offs. These works confirm that efficiency and accuracy must be considered jointly precisely the motivation for the composite metric proposed here yet none provides a unified multi-dimensional evaluation protocol.

2.2 Semantic Segmentation of UAV Imagery

Deep learning for UAV image segmentation has attracted substantial attention. UAVid established a dedicated benchmark for oblique drone video segmentation and highlighted the challenges of large scale variation and moving objects (Lyu et al., 2020), but its evaluation remains accuracy-centric. LoveDA exposed the domain gap between urban and rural remote sensing scenes (J. Wang et al., 2021), focusing on domain adaptation rather than deployment robustness. UNetFormer demonstrated that hybrid CNN–transformer designs can efficiently segment urban aerial scenes (L. Wang et al., 2022), yet reported no corruption, stability, or calibration analysis. A recent comprehensive review of UAV segmentation methods and datasets likewise reports evaluation practice centred on accuracy metrics (J. Cheng et al., 2024) and the study closest in spirit to ours a comparative evaluation of real-time segmentation networks on aerial flood imagery(Safavi & Rahnemoonfar, 2023) considers accuracy and speed but not robustness, calibration, stability, or spatial error structure. Across this literature, evaluation predominantly relies on mIoU and overall accuracy; computational efficiency is reported inconsistently, and robustness, calibration, and spatial error structure are typically omitted. Our work addresses this gap by evaluating all of these dimensions simultaneously on a common experimental base.

2.3 Robustness, Calibration, and Reliability

Robustness to common corruptions has been standardized for classification through ImageNet-C (Hendrycks & Dietterich, 2019) and extended to segmentation and detection in the driving context (Kamann & Rother, 2020; Michaelis et al., 2019). These studies show that architectural choices strongly influence corruption sensitivity, but comparable systematic analyses for UAV land-cover segmentation are scarce. Model calibration the agreement between predicted confidence and empirical accuracy was formalized for modern networks by (Guo et al., 2017) and connected to uncertainty estimation in dense prediction by (Kendall & Gal, 2017); deep ensembles were shown to improve both accuracy and uncertainty quality(Lakshminarayanan et al., 2016). Applications of these tools to drone segmentation remain limited, motivating the calibration, uncertainty, and ensemble analyses integrated in our framework.

2.4 Spatial Statistics in Error Analysis

Spatial autocorrelation statistics such as Moran's I (Moran, 1950) and its local variants (Anselin, 1995)are foundational in geography, epidemiology, and ecology for quantifying whether observations cluster in space. Although remote sensing accuracy assessment occasionally examines the spatial distribution of thematic errors,

to the best of our knowledge the direct application of Moran's I to deep segmentation error masks has not previously been reported. We adopt it as an analytical tool to test with permutation-based significance whether segmentation errors form spatial clusters, an insight with direct implications for boundary-aware post-processing such as conditional random fields (Krähenbühl & Koltun, 2012).

3. Proposed Methodology

Figure 1 provides an overview of the proposed framework: eight model configurations are trained under an identical protocol on the Semantic Drone Dataset and then assessed along four axes accuracy, efficiency, robustness/reliability, and spatial error structure culminating in scenario-dependent model recommendations.

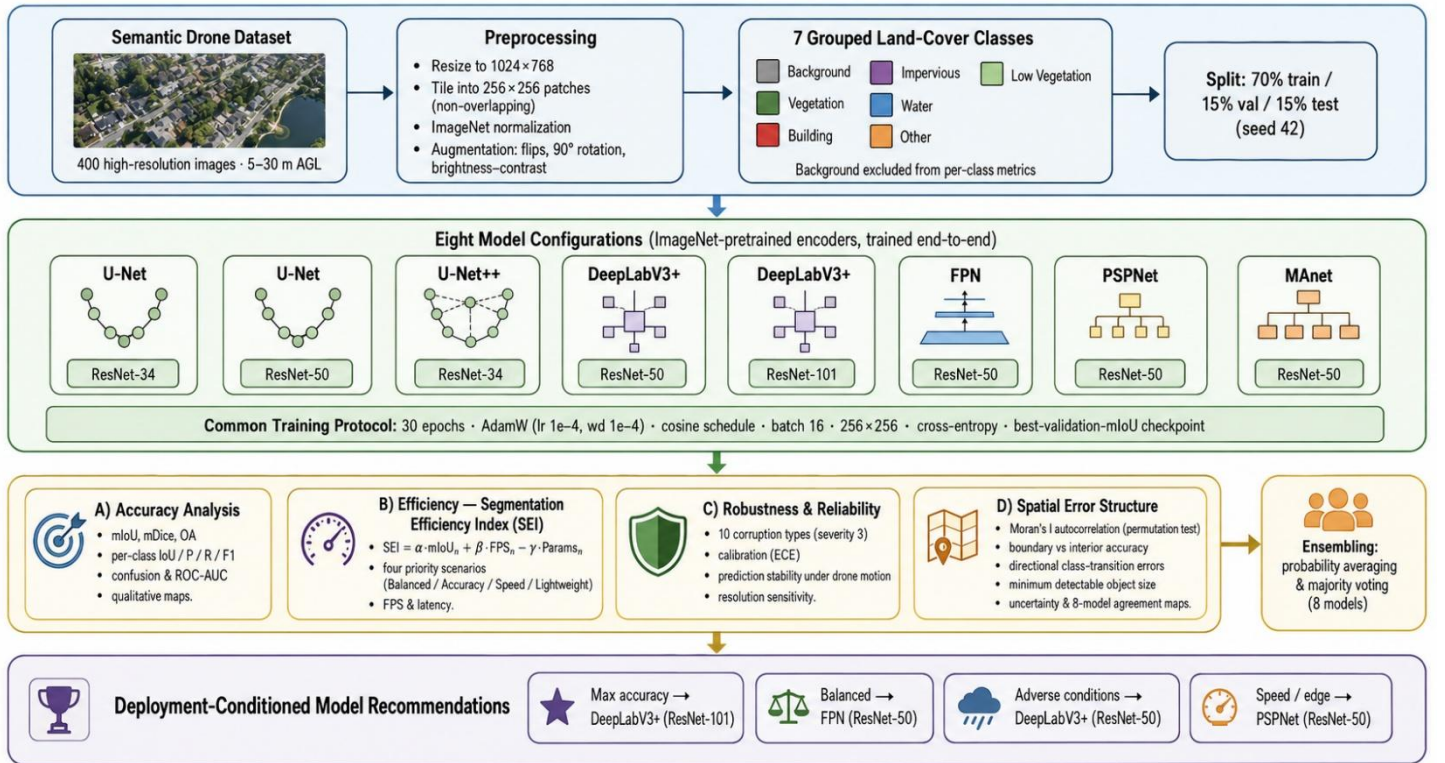


Figure 1: Workflow of the proposed deployment-oriented evaluation framework for UAV semantic segmentation: dataset preparation and class grouping, the eight model configurations trained under a common protocol, the four evaluation axes (accuracy, efficiency via the proposed SEI, robustness and reliability, and spatial error structure), and the resulting deployment-conditioned model recommendations (bottom row; derived in Sections 5–6).

3.1 Dataset Description

This study uses the Semantic Drone Dataset (*Institute of Computer Graphics and Vision (ICG), Graz University of Technology*, 2019), a publicly available benchmark of 400 high-resolution aerial images captured at 5–30 m above ground level over urban and suburban scenes. The original annotations comprise 24 fine-grained pixel-level classes. To reduce class imbalance and align with common practice in aerial land-cover analysis, the 24 classes are grouped into seven broader categories, as summarized in Table 1. The Background class combines unlabelled and conflicting areas, and as is common practice, it is included during training but not reported in per-class accuracy as it does not constitute a coherent semantic class.

Table 1. Mapping of the original 24 Semantic Drone Dataset classes to seven land cover classes.

Grouped Class	Original Classes
Background	unlabelled, inconsistent
Impervious	soil, gravel, and paved areas
Low Vegetation	Grass
Vegetation	Vegetation, tree, bald-tree
Water	pool, water
Building	Window, door, wall, roof
Other	car, bicycle, fence, fence-pole, pebbles, obstacle, ar-marker, person, dog other miscellaneous objects

The dataset is partitioned into three subsets: training (70%, 280 images), validation (15%, 60 images), and test (15%, 60 images) through stratified random sampling with a fixed seed of 42 to ensure reproducibility. This 70/15/15 partition adheres to the standard evaluation practice of balancing model fitting against reliable held-out estimation (Raschka, 2018). With a total of 400 images, the training would be weakened if the held-out sets were to be expanded, while reducing them to less than 60 images each would destabilise per-class statistics on rare classes. Each image is resized to 1024×768 pixels and tiled into non-overlapping 256×256 patches, yielding approximately 3,360 training, 720 validation, and 720 test patches after filtering predominantly background patches.

3.2 Data Preprocessing and Augmentation

All images are normalized with ImageNet channel statistics (mean 0.485, 0.456, 0.406); standard deviation (0.229, 0.224, 0.225), consistent with the pretrained encoders. Training augmentation combines horizontal flip ($p = 0.5$), vertical flip ($p = 0.5$), random 90° rotation ($p = 0.5$), and random brightness-contrast adjustment ($p = 0.3$). No augmentation is applied to the validation or test sets.

3.3 Representative Segmentation Architectures for Framework Validation

Eight model configurations spanning four architectural families are evaluated, all implemented with the Segmentation Models PyTorch (SMP) library (I. Iakubovskii, P., 2019) using ImageNet-pretrained encoders (He et al., 2016); we state the library explicitly because reported parameter counts depend on the implementation and encoder configuration. Table 2 summarizes the models; brief formulations follow.

Table 2. Summary of the evaluated architectures (all implemented in SMP with ImageNet-pretrained encoders).

Family	Model	Encoder	Params (M)	Key Mechanism
Encoder-decoder	U-Net	ResNet-34	24.4	Skip connections
Encoder-decoder	U-Net	ResNet-50	32.5	Skip connections
Encoder-decoder	U-Net++	ResNet-34	26.1	Nested dense skips
Atrous context	DeepLabV3+	ResNet-50	26.7	ASPP + decoder
Atrous context	DeepLabV3+	ResNet-101	45.7	ASPP + decoder
Pyramid	FPN	ResNet-50	26.1	Top-down feature pyramid

Pyramid	PSPNet	ResNet-50	24.3	Pyramid pooling (1,2,3,6)
Attention	MANet	ResNet-50	147.4	Multi-scale attention
				decoder

U-Net concatenates encoder features E_l with decoder features at matching resolutions, $D_l = f([Up(D_{l+1}); E_l])$, preserving fine spatial detail (Ronneberger et al., 2015); U-Net++ densifies these pathways with nested skips (Zhou et al., 2018). DeepLabV3+ applies atrous convolutions with rates r in parallel, $y[i] = \sum_k x[i + r \cdot k] \cdot w[k]$, aggregated by ASPP with global pooling (Chen et al., 2018). FPN fuses top-down pathways, $P_l = Up(P_{l+1}) + Conv_{\{1 \times 1\}}(C_l)$ (Lin et al., 2017). PSPNet concatenates pooled context at bin sizes $\{1, 2, 3, 6\}$ with the backbone feature map (Zhao et al., 2017). MANet inserts position-wise and multi-scale channel attention blocks in the decoder (Fan et al., 2020).

3.4 Proposed Segmentation Efficiency Index (SEI)

Current practice compares models by accuracy alone, ignoring the computational cost of achieving it. The SEI unifies accuracy, speed, and complexity into a single score:

$$SEI = \alpha \cdot mIoU_{norm} + \beta \cdot FPS_{norm} - \gamma \cdot Params_{norm}, \quad \alpha + \beta + \gamma = 1 \quad (1)$$

where $mIoU_{norm}$, FPS_{norm} , and $Params_{norm}$ are min–max normalized over the evaluated model pool, so SEI ranks models relative to the candidate set under consideration. The weights (α, β, γ) encode deployment priorities. Four scenarios are examined: SEI-Balanced (0.50, 0.30, 0.20), SEI-Accuracy (0.70, 0.20, 0.10), SEI-Speed (0.30, 0.50, 0.20), and SEI-Lightweight (0.30, 0.30, 0.40). Scores are reported on a 0 – 100 scale. Two properties of SEI should be stated explicitly. First, because normalization is performed over the candidate pool, SEI is a *relative* selection tool: adding or removing a model rescales all scores, which matches its intended use choosing among the finalists actually under consideration for a deployment rather than assigning absolute quality. Second, SEI complements rather than replaces Pareto-front analysis of the accuracy–speed–size trade-off: whereas the Pareto front identifies the set of non-dominated models, SEI provides a scalar ranking within and beyond that set once deployment priorities are stated, and a sensitivity analysis over the weights (Section 5.11) verifies that the reported rankings are stable over broad weight ranges rather than artifacts of a particular choice. The advantage of SEI over mIoU is that it exposes deployment-dependent rankings: a cloud pipeline would consult SEI-Accuracy, whereas an onboard edge processor would consult SEI-Speed or SEI-Lightweight.

3.5 Multi-Dimensional Evaluation Framework

3.5.1 Corruption Robustness Analysis

Robustness is evaluated under ten corruption types at severity level 3 of a five-level scheme following the design philosophy of ImageNet-C (Hendrycks & Dietterich, 2019); Table 3 specifies the exact parameterization of each corruption so the benchmark is reproducible. The corruption analysis is conducted on full test images resized to 256×256 (rather than the tiled-patch protocol of Section 3.1) to keep 10×8 corrupted evaluations tractable; clean-input mIoU under this protocol therefore differs from and is lower than the patch-level values of Section 5.1, and all robustness statistics are computed relative to the matching clean baseline. For each corruption type we measure the mIoU degradation relative to clean performance and aggregate a Robustness Score as the mean percentage drop across all ten corruptions.

Table 3. Parameterization of the ten corruption types (severity level 3).

Corruption	Parameterization (severity 3)
Gaussian noise	zero-mean, $\sigma = 0.08$ (images in $[0,1]$)
Motion blur	linear kernel, length 9 px, random angle
Fog	atmospheric blend, transmittance $t = 0.6$
Over-exposure	brightness gain +40%
Under-exposure	brightness gain -40%
Low contrast	contrast scale 0.5 about channel mean
Salt-and-pepper	impulse density 3% of pixels
JPEG compression	quality factor 15
Rain	additive streaks, 8% coverage, length 20 px
Shadow	random polygonal region, luminance $\times 0.5$

3.5.2 Model Calibration

Calibration is quantified by the Expected Calibration Error (ECE) (Guo et al., 2017). Pixel predictions are partitioned into $B = 15$ uniform-width confidence bins over $[0, 1]$, and the per-bin gap between accuracy and confidence is averaged with weights proportional to bin occupancy:

$$ECE = \sum_b (n_b / N) \cdot |acc(b) - conf(b)| \quad (2)$$

where n_b is the number of pixels in bin b , N the total pixel count, $acc(b)$ the empirical accuracy, and $conf(b)$ the mean predicted confidence in that bin (Guo et al., 2017). A perfectly calibrated model has $ECE = 0$. Uniform binning is used; we verified that adaptive (equal-mass) binning preserved the model ranking, and near-empty low-confidence bins are excluded from the reliability diagrams to avoid unstable per-bin estimates.

3.5.3 Spatial Error Autocorrelation (Moran's I)

To test whether errors cluster spatially, we compute Moran's I (Moran, 1950) on a binary error mask, where $x_i = 1$ if the predicted label of pixel i differs from the ground truth and $x_i = 0$ otherwise:

$$I = (n / W) \cdot [\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})] / \sum_i (x_i - \bar{x})^2 \quad (3)$$

with $w_{ij} = 1$ for 8-connected neighboring pixels (0 otherwise), $W = \sum_i \sum_j w_{ij}$, and n the number of pixels; masks are computed on 64×64 downsampled error maps for tractability. I ranges from -1 (dispersed) through ≈ 0 (random) to $+1$ (clustered); under the null hypothesis of spatial randomness the expected value is $E[I] = -1/(n - 1) \approx 0$. Statistical significance is assessed with a 999-iteration random permutation test, reporting pseudo p-values, and per-model dispersion is reported as ± 1 standard deviation of I across test images. Error density maps are additionally generated by Gaussian kernel density estimation (bandwidth 15 px) over error pixel locations to visualize clustering hotspots.

3.5.4 Prediction Stability Analysis

Stability under drone motion is measured by applying geometric perturbations to test images: rotation, horizontal shift, vertical shift, and zoom. Each perturbation type is applied individually (never composited); for each test image and each type, 10 random perturbations are sampled uniformly rotation from $[-5^\circ, +5^\circ]$, shifts from $[-10, +10]$ px, and zoom factor from $[0.95, 1.05]$ and results are averaged over all samples and seeds. The pixel flip rate is a metric that quantifies stability. It is the percentage of pixels whose predicted class changes between the original and the (inverse-warped) perturbed prediction, as assessed on the valid overlapping region.

3.5.5 Class Transition Error Analysis

The analysis of boundary behaviour is conducted on boundary pixels, which are operationally defined as pixels that are located within a 3-pixel band of a ground-truth class edge that is obtained by morphological gradient of the label map. This definition is consistent with the common practice of evaluating boundary-quality in segmentation. In contrast to a global confusion matrix, this analysis is direction-aware and limited to the locations where two ground-truth classes intersect: the error rate of transition. A to B (pixels of class) The directional biases in the manner in which models resolve boundary ambiguity are revealed by the distinction between $B \rightarrow A$ and A adjacent to class B. In order to prevent unstable estimates, transitions supported by fewer than 500 boundary pixels are excluded from the tabulation, and the supporting pixel count of each transition is reported in conjunction with its error rate, as some class pairs share very brief boundaries.

3.5.6 Model Prediction Diversity

The diversity of each model pair (m, m') is determined by the fraction of test pixels for which the two predicted labels differ. This is how ensemble potential is evaluated: through pairwise prediction disagreement. Higher disagreement between individually accurate models indicates complementary error patterns and thus greater expected ensemble benefit (Lakshminarayanan et al., 2016).

3.5.7 Minimum Detectable Object Size

For each class, ground-truth connected components are binned by pixel area, and the detection rate of a bin is defined as the fraction of ground-truth instances for which at least 50% of the instance's pixels receive the correct class label. This yields, per class and model, the minimum object area at which detection becomes reliable directly informing flight-altitude and sensor-resolution planning. Bins supported by very few instances (most notably for the spatially contiguous Water and Low Vegetation classes, which form few connected components) are sparsely populated, and the corresponding curves are interpreted with caution.

3.5.8 Uncertainty, Agreement, Boundary Accuracy, and Ensembling

Pixel-level uncertainty maps are computed from the prediction entropy of the SoftMax output (Kendall & Gal, 2017). Multi-model agreement maps record, per pixel, the number of models (out of eight) voting for the majority label. Boundary versus interior accuracy contrasts pixel accuracy inside the 3-pixel boundary band (Section 3.5.5) with accuracy elsewhere. Finally, two ensembling strategies are compared: (i) probability averaging, in which the eight SoftMax probability maps are averaged before the argmax decision, and (ii) majority voting, in which each pixel receives the label predicted by the largest number of individual models, with ties resolved by the mean probability.

4. Experimental Setup

4.1 Implementation Details

All experiments were conducted with PyTorch 2.5.1 and CUDA 12.4 (Paszke et al., 2019) on an NVIDIA GeForce RTX 4080 SUPER GPU (16 GB). Models were implemented with Segmentation Models PyTorch version 0.3.4 (Loshchilov & Hutter, 2017) using ImageNet-pretrained encoders. Training ran for 30 epochs with the AdamW optimizer (P. Iakubovskii, 2019), initial learning rate 1×10^{-4} , weight decay 1×10^{-4} , cosine annealing schedule, batch size 16, input resolution 256×256 , and cross-entropy loss; the checkpoint with the highest validation mIoU was retained. Each configuration was trained once under this fixed-seed protocol; the implications of single-run evaluation for the interpretation of small accuracy margins are discussed in Sections 5.1 and 6.3. Under this budget, per-model training required between 5.0 minutes (PSPNet, ResNet-50) and 12.9 minutes (MAnet, ResNet-50), approximately 69 GPU-minutes in total, and best validation mIoU ranged from 0.6859 (PSPNet) to 0.7324 (U-Net, ResNet-50). Validation mIoU plateaued for all models within the final five epochs, indicating convergence under this budget given ImageNet pretraining and the patch-level dataset size; the effect of substantially longer schedules is acknowledged as a limitation in Section 6.3. Inference speed (FPS) is measured on the same GPU with batch size 1 at 256×256 after 50 warm-up iterations, averaged over 500 runs, using the model forward pass only (excluding data loading and post-processing).

4.2 Evaluation Metrics

The metric suite comprises mean Intersection over Union (mIoU), mean Dice coefficient (mDice), Overall Accuracy (OA), per-class IoU, precision, recall, F1-score, Area Under the ROC Curve (AUC), Expected Calibration Error (ECE, Eq. 2), inference speed (FPS) and latency, parameter count, and the proposed SEI (Eq. 1). For a class c with true positives TP_c , false positives FP_c , and false negatives FN_c , $IoU_c = TP_c / (TP_c + FP_c + FN_c)$ and $Dice_c = 2TP_c / (2TP_c + FP_c + FN_c)$; mIoU and mDice average these over the six semantic classes (excluding Background, Section 3.1). It is important to note that the Dice coefficient is mathematically identical to the F1-score for per-class binary evaluation (both are equivalent to the harmonic mean of precision and recall). Consequently, a single value is reported per class; mDice is therefore equivalent to the macro F1-score.

5. Results and Discussion

The core evidence for each evaluation dimension is presented in the main-text figures, while Supplementary Figures S1–S13 provide extended visualisations, which are referenced as necessary.

5.1 Multi-Architecture Comparison

The test set's overall comparison is presented in Table 4. In accordance with the advantages of a deeper encoder and atrous spatial pyramid pooling for multi-scale aerial content, DeepLabV3+ with ResNet-101 obtains the highest mIoU (0.7172), mDice (0.7811), and overall accuracy (92.33%). FPN with ResNet-50 achieves a competitive mIoU of 0.7117, which is within 0.8% of the best, despite having significantly fewer parameters (26.1M vs. 45.7M) and a 68% higher throughput (292.4 vs. 174.4 FPS). This suggests an exceptional accuracy–efficiency trade-off. The lowest mIoU (0.6785) is achieved by PSPNet with ResNet-50, which is by far the quickest model (622.8 FPS). This is likely due to the fact that aggressive global pooling discards fine spatial detail that is essential for pixel-level prediction. It is important to note that MAnet, the largest model with 147.4M parameters, is the slowest (165.1 FPS) and only the seventh most accurate. This demonstrates that model capacity alone does not guarantee superior performance; this observation directly motivates the SEI metric introduced in Section 3.4. Two caveats frame the interpretation of these rankings. First, models two through six span only 0.6 mIoU points (0.7060–0.7117); as each configuration was trained once (Section 4.1), such sub-point margins are plausibly within run-to-run variance, and we accordingly refrain from strong ordinal claims inside this cluster, focusing instead on the large and consistent separations the ~4-point gap to PSPNet, the ~1.8-point gap between the leader and MAnet, and the pronounced differences in the efficiency, robustness, calibration, and stability dimensions analysed below. Second, within that cluster U-Net gains little from encoder depth: the ResNet-50 variant (0.7093) is essentially tied with DeepLabV3+ (ResNet-50) while U-Net (ResNet-34) trails by only 0.3 points at 28% higher speed.

Table 4. Overall comparison of the eight model configurations on the Semantic Drone Dataset test set (best values in bold)

Model	Params (M)	mIoU	mDice	OA (%)	FPS	Latency (ms)
DeepLabV3+ (ResNet-101)	45.7	0.7172	0.7811	92.33	174.4	5.7
FPN (ResNet-50)	26.1	0.7117	0.7788	92.03	292.4	3.4
U-Net (ResNet-50)	32.5	0.7093	0.7741	92.11	299.6	3.3
DeepLabV3+ (ResNet-50)	26.7	0.7092	0.7763	92.01	303.2	3.3
U-Net++ (ResNet-34)	26.1	0.7086	0.7734	91.94	287.0	3.5
U-Net (ResNet-34)	24.4	0.7060	0.7715	91.74	383.3	2.6
MAnet (ResNet-50)	147.4	0.6990	0.7671	91.38	165.1	6.1
PSPNet (ResNet-50)	24.3	0.6785	0.7557	90.15	622.8	1.6

Figure 2 presents qualitative comparisons on five representative test samples, and Figure 3 summarizes the quantitative results, contrasting mIoU and mDice across models and reporting per-class IoU for the best model. The complexity–performance relationship (Supplementary Fig. S1) confirms that accuracy is not monotone in parameter count DeepLabV3+ (ResNet-101) leads despite MAnet being three times larger demonstrating that architectural design outweighs raw capacity, with FPN occupying the most attractive region of the accuracy complexity plane.

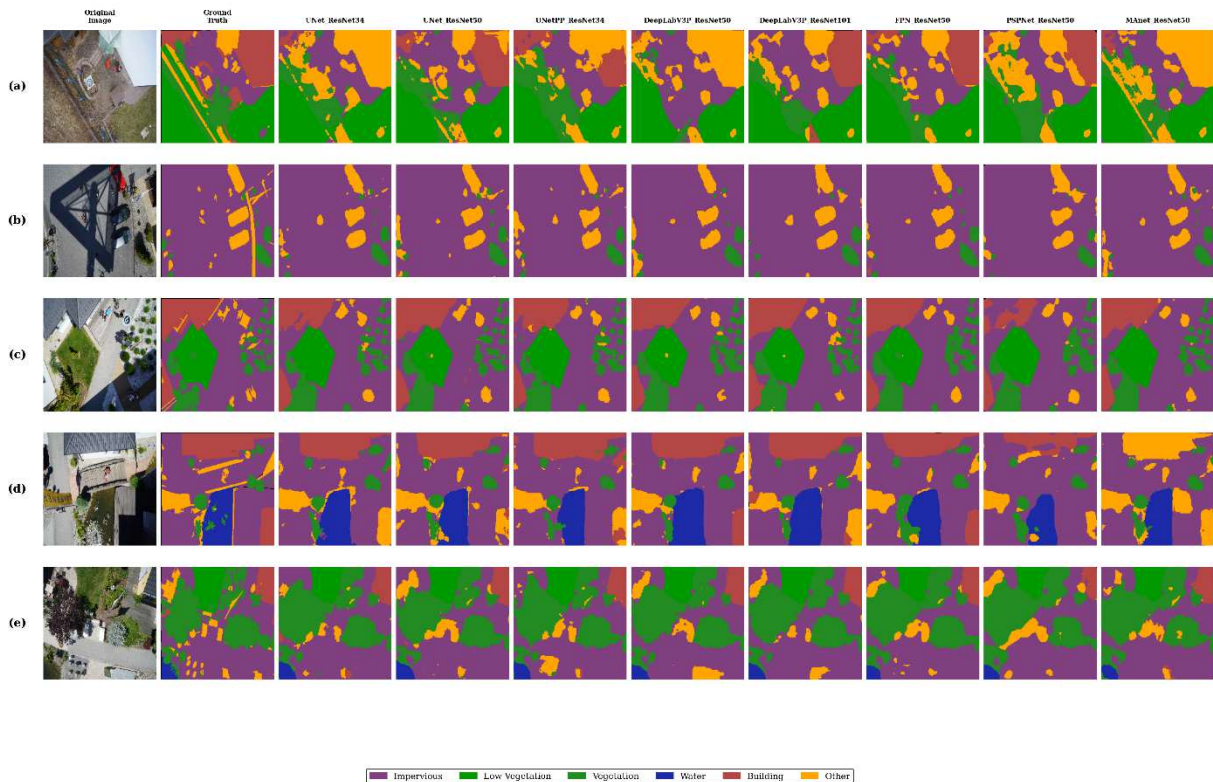


Figure 2: Qualitative comparison of segmentation results across all eight model configurations on five test samples (rows a–e). Classes: Impervious (purple), Low Vegetation (light green), Vegetation (dark green), Water (blue), Building (red), Other (orange).

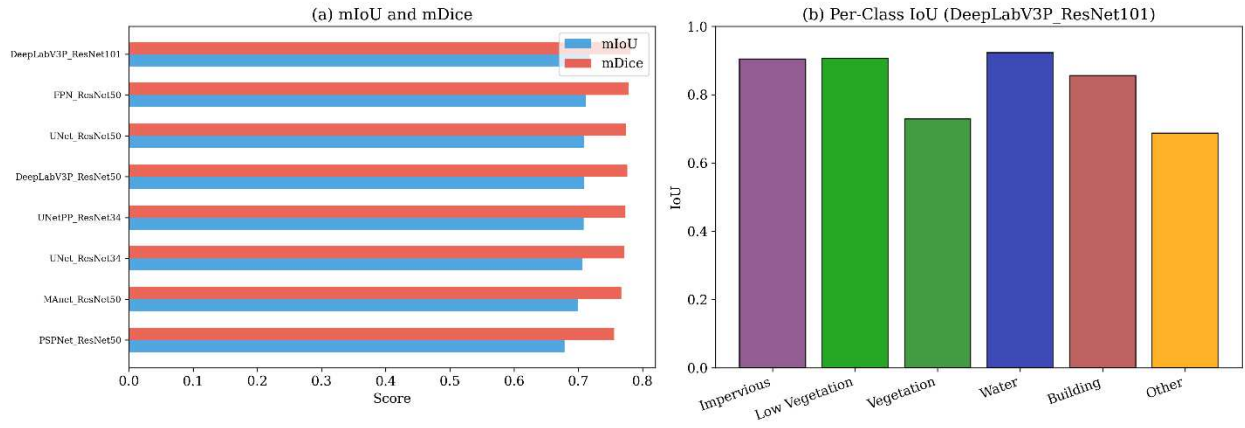


Figure 3: Quantitative performance analysis. (a) mIoU and mDice across all models. (b) Per-class IoU for the best model, DeepLabV3+ (ResNet-101).

5.2 Per-Class Performance and Confusion Analysis

The top model's per-class metrics are delineated in Table 5. The difficulty of segmentation differs significantly among the six semantic land-cover categories (Background is excluded from per-class reporting as outlined in Section 3.1). Water has the highest IoU (0.9231) due to its unique appearance and spatial uniformity. Low vegetation (0.9069) and impervious surfaces (0.9044) also exhibit strong performance, owing to their consistent texture and extensive spatial extent. In contrast, the heterogeneous Other class has the lowest IoU (0.6875), which is indicative of its composition of small objects (vehicles, persons, fences). Vegetation (0.7289) is frequently confused with Low Vegetation and Impervious regions, while Building (0.8563) is challenged by a variety of roof materials and orientations.

Table 5. Per-class metrics for the most effective model, DeepLabV3+ (ResNet-101) (the most advantageous values are depicted in bold). The per-class Dice coefficient is equivalent to F1 (Section 4.2).

Class	IoU	Precision	Recall	F1 / Dice
Impervious	0.9044	0.9334	0.9668	0.9498
Low Vegetation	0.9069	0.9628	0.9398	0.9512
Vegetation	0.7289	0.8536	0.8331	0.8432
Water	0.9231	0.9574	0.9626	0.9600
Building	0.8563	0.9369	0.9087	0.9226
Other	0.6875	0.8354	0.7952	0.8148

The best model's normalised confusion matrix (Figure 4) reveals systematic misclassification patterns among the six semantic classes. The largest off-diagonal mass in the Vegetation row flows to Impervious (0.11), reflecting a shadowed canopy over paved ground. A secondary Vegetation→Low Vegetation confusion (0.03) is driven by similar spectral characteristics; the heterogeneous Building pixels are occasionally designated Impervious (0.05), which is consistent with the visual similarity of certain roofing materials and paved surfaces. The other class disperses across multiple categories, with the strongest dispersion toward Impervious (0.11). Per Section 3.1, the Background row is constructed to be diffuse and is therefore excluded from all per-class statistics.

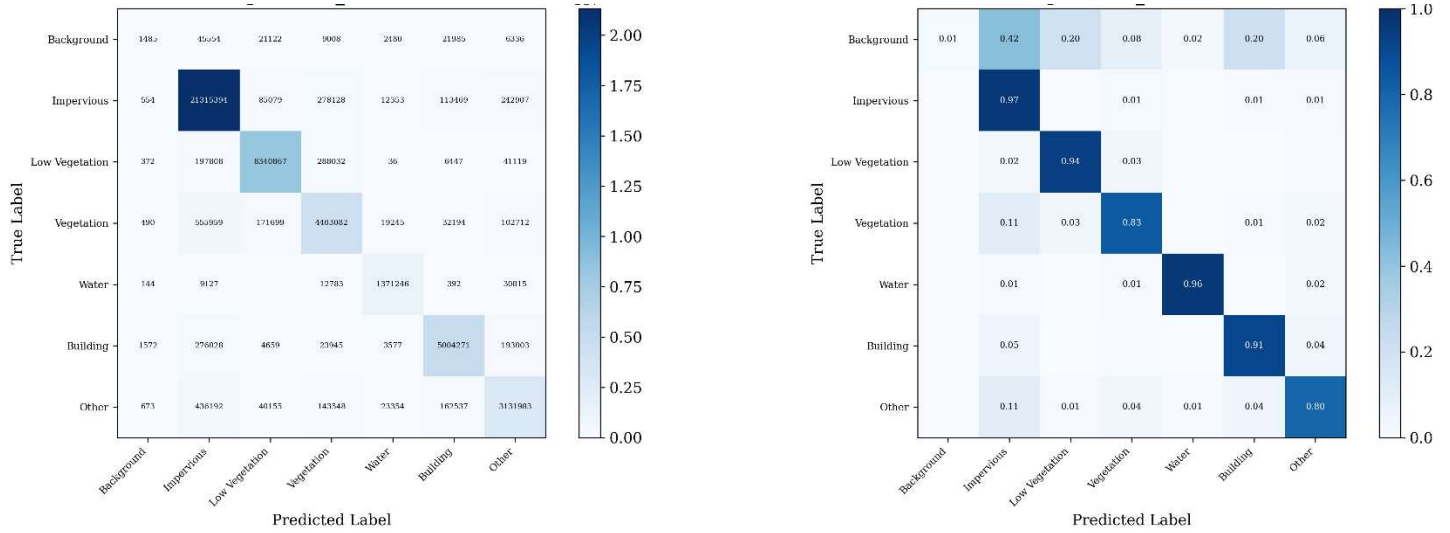


Figure 4: Confusion matrix analysis for DeepLabV3+ (ResNet-101). (a) Raw counts. (b) Row-normalized proportions.

5.3 Precision, Recall, and F1-Score Evaluation

Macro-averaged precision, recall, and F1 (Supplementary Fig. S2) mirror the mIoU ordering of Table 4, with DeepLabV3+ (ResNet-101) achieving the highest macro-F1 (0.7811) and FPN close behind (0.7788); precision consistently exceeds recall for all models (e.g., 0.8229 vs. 0.7743 for the best model), indicating a systematic tendency toward conservative labelling in which minority-class pixels are under-retrieved rather than over-predicted. The per-class F1 heatmap across all models (Supplementary Fig. S3) shows that class difficulty patterns are consistent across architectures rather than model-specific: every model scores 0.93–0.95 on Impervious and Low Vegetation but only 0.73–0.82 on Other, where PSPNet is weakest (0.73). One-vs-rest ROC analysis for the best model (Supplementary Fig. S4) confirms high pixel-level score separability for every class AUC value of 0.986–1.000, lowest for the heterogeneous Other class indicating that the practical difficulty of hard classes lies in the argmax competition with co-occurring classes rather than in raw score separability.

5.4 Inference Speed Analysis

Table 6 and Figure 5 quantify the accuracy–speed trade-off. PSPNet achieves the highest throughput (622.8 FPS; 1.6 ms latency), 3.6× faster than DeepLabV3+ (ResNet-101) (174.4 FPS; 5.7 ms), making it the natural high-throughput candidate despite its lower accuracy (mIoU 0.6785 vs. 0.7172, a 5.4% relative gap). FPN offers an attractive middle ground: within 0.8% of the best mIoU while sustaining 292.4 FPS. On this desktop-class GPU at 256×256 resolution, every model comfortably exceeds a 30 FPS real-time threshold; the throughput differences instead determine the headroom available for larger inputs, batched multi-stream processing, or migration to far weaker embedded hardware, where the nearly fourfold spread between PSPNet and MANet (165.1 FPS, the slowest model) becomes decisive.

Table 6. Inference speed comparison (batch size 1, 256×256 input, RTX 4080 SUPER, forward pass only; best values in bold).

Model	FPS	Latency (ms)	Params (M)
PSPNet (ResNet-50)	622.8	1.6	24.3
U-Net (ResNet-34)	383.3	2.6	24.4
DeepLabV3+ (ResNet-50)	303.2	3.3	26.7
U-Net (ResNet-50)	299.6	3.3	32.5
FPN (ResNet-50)	292.4	3.4	26.1
U-Net++ (ResNet-34)	287.0	3.5	26.1
DeepLabV3+ (ResNet-101)	174.4	5.7	45.7
MAnet (ResNet-50)	165.1	6.1	147.4

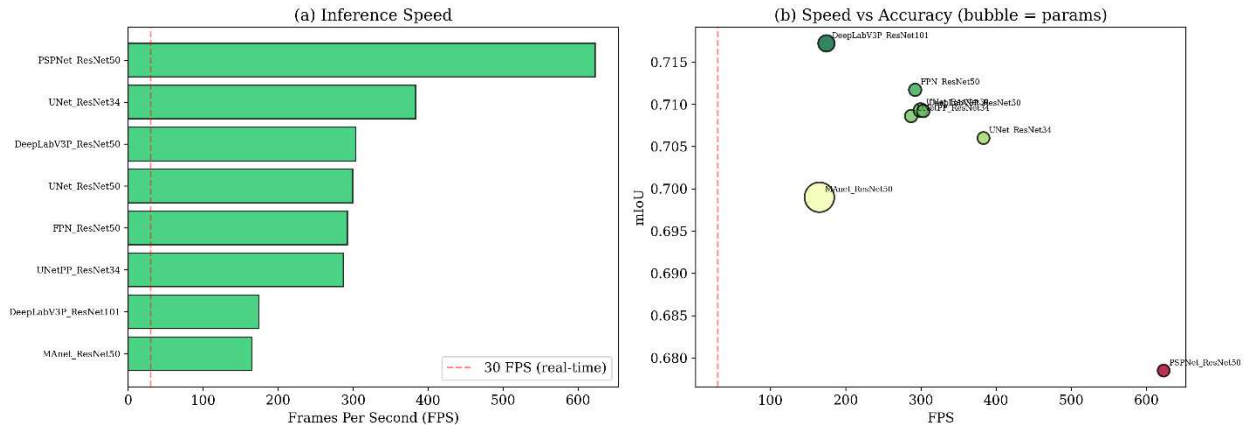


Figure 5: Inference speed analysis. (a) FPS across models. (b) Speed–accuracy trade-off with the 30 FPS real-time threshold indicated (bubble size = parameters).

5.5 Visual Error, Uncertainty, and Multi-Model Agreement Analysis

Qualitative error maps (Supplementary Fig. S5) corroborate the quantitative findings: correct predictions dominate the interiors of large regions, while errors concentrate along class boundaries, on small objects, and in regions of unusual appearance such as shadows and occlusions exactly the spatial structure identified quantitatively by the boundary and Moran's I analyses (Sections 5.6 and 5.10). Entropy-based uncertainty maps (Supplementary Fig. S6) show that model uncertainty concentrates at the same locations, while interiors of large homogeneous regions are confidently predicted. The multi-model agreement analysis (Supplementary Fig. S7) is complementary: pixels where all eight models agree are highly reliable, whereas disagreement flags regions warranting human verification or additional data collection. Together these maps constitute a practical quality-assurance layer for automated drone mapping.

5.6 Boundary vs. Interior Accuracy

Table 7 and Figure 6 quantify a phenomenon often observed qualitatively but rarely measured: interior pixels are far more accurately classified than boundary pixels for every model, with accuracy gaps of approximately 41–43 percentage points under the 3-pixel boundary-band definition of Section 3.5.5. U-Net (ResNet-50) attains the best boundary accuracy (50.91%) and the smallest gap (41.43 points), consistent with the boundary-preserving role of

its full-resolution skip connections, while DeepLabV3+ (ResNet-101) achieves the best interior accuracy (92.72%); PSPNet is weakest on both counts (47.87% boundary accuracy; 42.88-point gap), reflecting the localization cost of aggressive global pooling. The magnitude of this gap underscores the importance of boundary-aware loss functions and boundary-targeted post-processing for improving overall quality.

Table 7. Boundary vs. interior pixel accuracy (%) and interior–boundary gap (best values in bold).

Model	Boundary Accuracy	Interior Accuracy	Gap
U-Net (ResNet-50)	50.91	92.34	41.43
U-Net++ (ResNet-34)	50.73	92.31	41.58
DeepLabV3+ (ResNet-101)	50.55	92.72	42.17
DeepLabV3+ (ResNet-50)	50.30	92.28	41.99
U-Net (ResNet-34)	50.00	92.06	42.06
FPN (ResNet-50)	49.97	92.44	42.46
MAnet (ResNet-50)	49.32	91.58	42.26
PSPNet (ResNet-50)	47.87	90.75	42.88

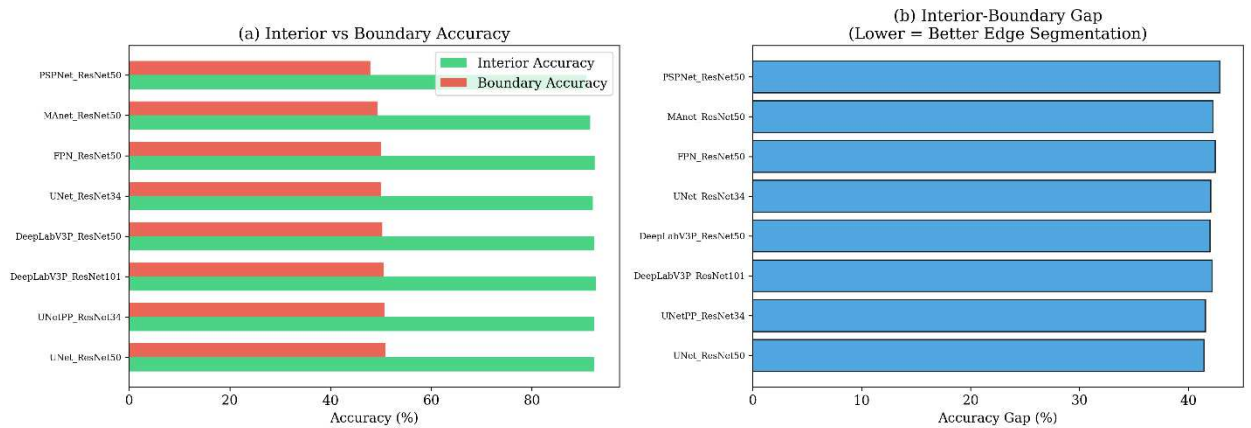


Figure 6: Boundary vs. interior accuracy. (a) Per-model comparison. (b) Interior–boundary gap (lower = better edge segmentation).

5.7 Ensemble Analysis

Comparing the two ensembling strategies of Section 3.5.8 against all individual models (Supplementary Fig. S8), averaging the softmax probabilities of all eight models yields an ensemble mIoU of 0.7223 and OA of 92.66% a 0.7% relative (+0.51 mIoU points) improvement over the best single model (0.7172) while majority voting reaches 0.7210 (+0.5% relative), confirming that soft probability fusion preserves more information than hard label voting. Although modest in absolute terms, the gain is obtained without any retraining and outperforms every individual architecture; the per-class comparison shows it concentrates in the challenging classes (Other, Vegetation) where individual models are most variable, confirming that architectural diversity (Section 5.12) translates into complementary corrections. Aggregating per-class IoU across all eight architectures (Supplementary Fig. S9) supports the same conclusion from the difficulty side: Water, Low Vegetation, and Impervious display high medians with low inter-model variance, whereas Vegetation and especially other show

lower medians and wider spreads (PSPNet a low outlier at 0.5697 on Other) the intrinsically hard classes where both architectural choice and ensembling matter most.

5.8 Segmentation Efficiency Index (SEI) Results

SEI scores computed with Eq. (1) under the four priority scenarios of Section 3.4 are reported in Table 8 and Figure 7. FPN (ResNet-50) achieves the highest score (50.95) under SEI-Balanced, despite not having the best mIoU, due to its favourable joint profile of accuracy, speed, and compactness. DeepLabV3+ (ResNet-101), despite leading in raw mIoU and dominating SEI-Accuracy (68.67), ranks only fifth under balanced priorities once its parameter count and slower inference are taken into account. The rankings are significantly affected by the different scenarios: PSPNet is the dominant model in SEI-Speed (50.00), while U-Net (ResNet-34) is the top performer in SEI-Lightweight (35.59) due to its high speed, respectable accuracy, and nearly minimal footprint. MAnet is penalised in all scenarios due to the fact that its SEI-Speed and SEI-Lightweight scores are negative, as it aligns the highest parameter count with the lowest throughput. These scenario-dependent reversals are the deployment-critical nuance that a single mIoU number cannot convey.

Table 8. Under four priority scenarios, SEI scores (0–100 scale) are presented (the highest value for each scenario is highlighted in bold).

Model	mIoU	FPS	SEI-Bal	SEI-Acc	SEI-Speed	SEI-Light
FPN (ResNet-50)	0.7117	292.4	50.95	65.47	39.35	33.50
U-Net (ResNet-34)	0.7060	383.3	49.82	59.27	45.14	35.59
DeepLabV3+ (ResNet-50)	0.7092	303.2	48.33	61.37	38.49	32.07
U-Net (ResNet-50)	0.7093	299.6	47.28	60.92	37.24	30.03
DeepLabV3+ (ResNet-101)	0.7172	174.4	47.13	68.67	27.54	23.66
U-Net++ (ResNet-34)	0.7086	287.0	46.59	59.62	36.36	30.74
PSPNet (ResNet-50)	0.6785	622.8	30.00	20.00	50.00	30.00
MAnet (ResNet-50)	0.6990	165.1	6.49	27.08	−4.11	−24.11

To verify that these conclusions are not artifacts of a particular weight choice, Table 9 reports a weight-sensitivity analysis computed directly from the quantities underlying Table 8: the accuracy weight α is swept over $[0.30, 0.90]$ while the speed-to-size ratio $\beta:\gamma$ is held at 3:2 and $\beta + \gamma = 1 - \alpha$. The top-ranked model transitions smoothly and interpretably along the priority axis PSPNet for strongly speed-dominated weightings ($\alpha \leq 0.35$), U-Net (ResNet-34) in a narrow intermediate band, FPN across the broad balanced region ($\alpha \approx 0.45\text{--}0.65$, which contains the SEI-Balanced setting), and DeepLabV3+ (ResNet-101) once accuracy dominates ($\alpha \geq 0.70$). FPN's balanced-scenario leadership is therefore stable over a wide neighbourhood of weights rather than specific to (0.50, 0.30, 0.20).

Table 9. SEI weight-sensitivity analysis: top-ranked model as the accuracy weight α varies ($\beta:\gamma$ fixed at 3:2; computed from the Table 8 inputs).

Accuracy weight α	Top-ranked model	Interpretation
0.30–0.35	PSPNet (ResNet-50)	Speed-dominated selection
0.40–0.45	U-Net (ResNet-34)	Speed–accuracy compromise
0.45–0.65	FPN (ResNet-50)	Balanced region (incl. SEI-Balanced)
0.70–0.90	DeepLabV3+ (ResNet-101)	Accuracy-dominated selection

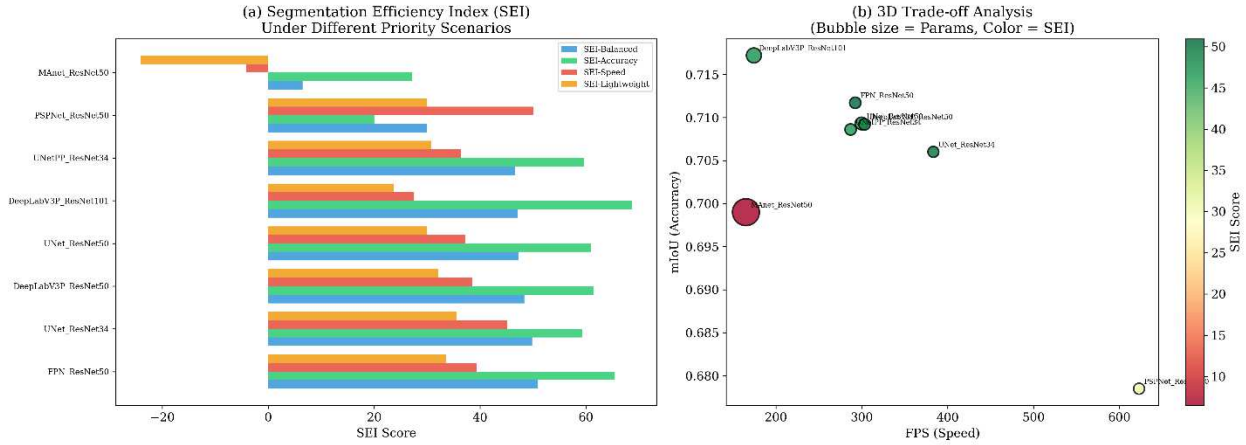


Figure 7: Segmentation Efficiency Index results. (a) SEI scores under the four priority scenarios. (b) Accuracy–speed–size trade-off (bubble size = parameters; color = SEI-Balanced score).

5.9 Corruption Robustness Analysis

Table 10 and Figure 8 summarize robustness under the ten corruptions of Table 3, evaluated with the resized-image protocol of Section 3.5.1. Average performance degrades by 20–27% relative to clean inputs, with pronounced architectural variation: DeepLabV3+ (ResNet-50) is the most robust (20.4% mean drop), suggesting that atrous multi-scale context aggregation confers inherent resilience, whereas MAenet (26.8%) and PSPNet (26.7%) degrade most. Interestingly, robustness follows neither accuracy nor encoder depth the ResNet-50 DeepLabV3+ variant is more resilient than its ResNet-101 counterpart (23.2%), and the least accurate model (PSPNet) is also among the least robust. Across corruption types, the noise-family degradations are by far the most damaging: Gaussian noise (mIoU 0.19–0.31 across models) and salt-and-pepper noise (0.16–0.34) roughly halve performance, followed by motion blur and JPEG compression, whereas photometric perturbations under-exposure, shadow, and low contrast have comparatively mild impact. Visual failure modes (Supplementary Fig. S10) illustrate the mechanisms: heavy noise collapses predictions into large erroneous regions, motion blur erases small objects, and fog merges spectrally similar classes. These findings carry a direct operational implication mission anticipating specific degradations should select architectures with demonstrated resilience to those corruption types, and training pipelines should prioritize augmentation with the most damaging (noise-type) corruptions.

Table 10. Corruption robustness under the resized-image protocol: clean and average corrupted mIoU with mean relative drop (best values in bold).

Model	Clean mIoU	Avg. Corrupted mIoU	Mean Drop (%)	Rank
DeepLabV3+ (ResNet-50)	0.5988	0.4767	20.40	1
U-Net++ (ResNet-34)	0.6059	0.4698	22.46	2
FPN (ResNet-50)	0.6068	0.4705	22.47	3
DeepLabV3+ (ResNet-101)	0.6281	0.4821	23.25	4
U-Net (ResNet-50)	0.6150	0.4616	24.95	5
U-Net (ResNet-34)	0.6005	0.4504	25.00	6
PSPNet (ResNet-50)	0.5324	0.3905	26.65	7
MAnet (ResNet-50)	0.5805	0.4248	26.81	8

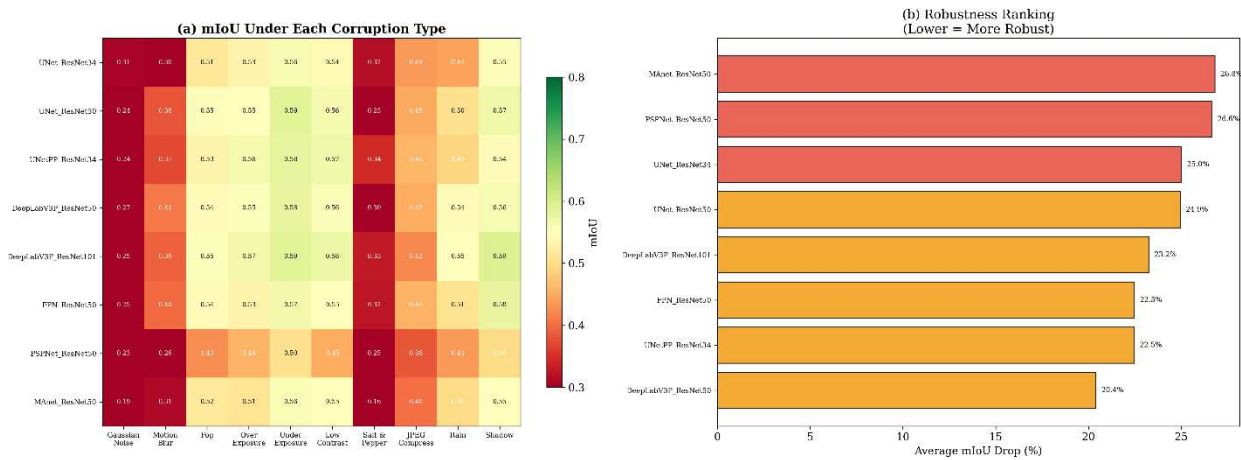


Figure 8: Corruption robustness analysis. (a) mIoU heatmap across the ten corruption types. (b) Robustness ranking by mean mIoU drop (lower = more robust).

5.10 Model Calibration Analysis

The ECE analysis (Figure 9) shows that all models are reasonably well calibrated, with ECE below 0.07. PSPNet (ResNet-50) is the best-calibrated model (ECE = 0.0397), followed by U-Net++ (ResNet-34, 0.0457) and U-Net (ResNet-34, 0.0479), while FPN is the least calibrated (0.0637). It is crucial to note that the most accurate model, DeepLabV3+ (ResNet-101), is mid-pack (0.0493), while the least accurate model is the best calibrated, echoing the decoupling reported for classification networks (Guo et al., 2017). In other words, increased accuracy does not necessarily imply better calibration. The reliability curves are consistently below the diagonal, indicating that all models are systematically overconfident. This is critical for safety-critical applications, as a model that is not properly calibrated may attribute a high degree of confidence to erroneous predictions, which could lead to misleading downstream decisions in autonomous navigation or disaster assessment.

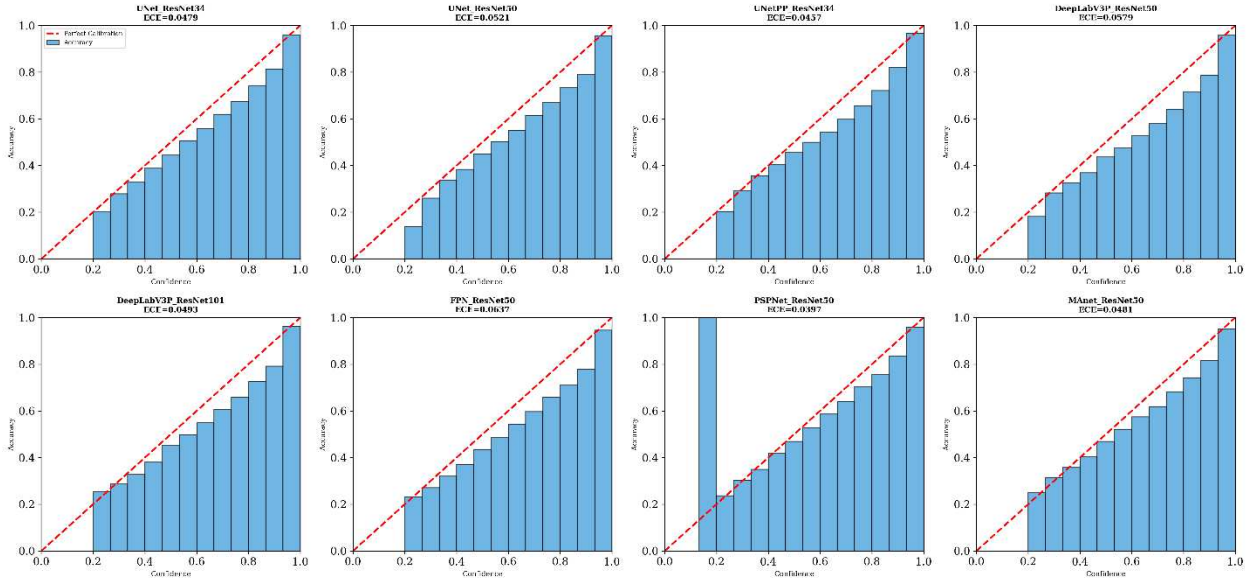


Figure 9: Reliability diagrams and ECE values for all eight models (15 uniform confidence bins; near-empty bins excluded).

5.11 Prediction Stability and Resolution Sensitivity

Under the perturbation protocol of Section 3.5.4, the mean pixel flip rates vary across architectures, ranging from 9.4% to 11.2% (Figure 10). DeepLabV3+ models are the most stable — 9.36% for the ResNet-50 variant and 9.48% for ResNet-101 — plausibly because dilated convolutions yield smoother, more translation-tolerant feature responses, whereas U-Net (ResNet-50) is the least stable (11.24%). Across perturbation types, horizontal shift and rotation cause the greatest instability (up to 12.5% flips), while zoom is the mildest (7.5–9.6%). Stability is critical for video-based drone applications, where temporal flicker in segmentation undermines downstream tracking and mapping; notably, the ~10% of pixels that flip under even sub-degree, few-pixel perturbations coincide largely with the boundary bands identified in Sections 5.6 and 5.12, indicating that instability and boundary error share a common spatial origin. The related resolution-sensitivity analysis (Supplementary Fig. S11) shows all models improving steeply from 128 to 384 px input resolution, with DeepLabV3+ variants degrading most gracefully at low resolution (0.4662 mIoU at 128 px for ResNet-101), PSPNet the most resolution-hungry, U-Net++ the least sensitive overall (mIoU range 0.190, saturating with a marginal $0.6208 \rightarrow 0.6181$ decline from 384 to 512 px), and U-Net (ResNet-34) the most sensitive (range 0.225).

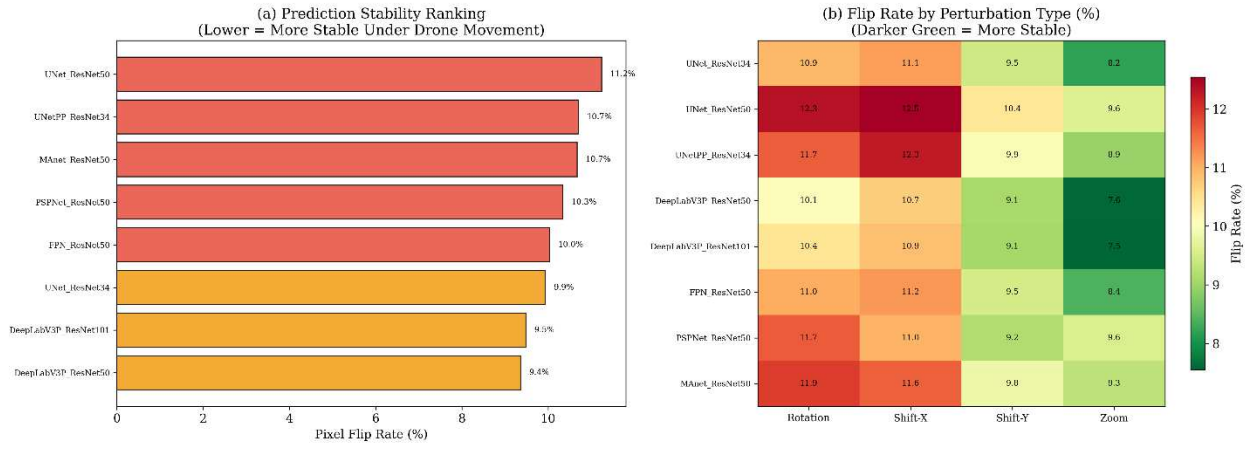


Figure 10: Prediction stability under simulated drone movement. (a) Stability ranking by mean pixel flip rate (lower = more stable). (b) Flip-rate heatmap by perturbation type.

5.12 Spatial Error Autocorrelation (Moran's I)

Moran's I computed on binary error masks (Eq. 3) is strongly and significantly positive for all models ranging from 0.797 ± 0.092 for DeepLabV3+ (ResNet-50) to 0.857 ± 0.041 for PSPNet (permutation test, 999 iterations, pseudo $p < 0.001$ for every model and image) confirming that segmentation errors are spatially clustered rather than random (Figure 11). It is notable that the model ordering inversely tracks accuracy and robustness: PSPNet, the weakest model, produces the most spatially clustered errors, indicating that its mistakes form large contiguous blobs rather than scattered pixels. These clusters are primarily located along class boundaries and within visually ambiguous regions, as indicated by the kernel-density error maps. The practical implications are significant: targeted post-processing dense conditional random fields (Krähenbühl & Koltun, 2012) or boundary refinement modules can be selectively applied to error-prone regions, resulting in accuracy gains with minimal computational overhead, as errors concentrate in predictable spatial structures. This is the first reported application of Moran's I to deep segmentation error analysis, to the best of our knowledge.

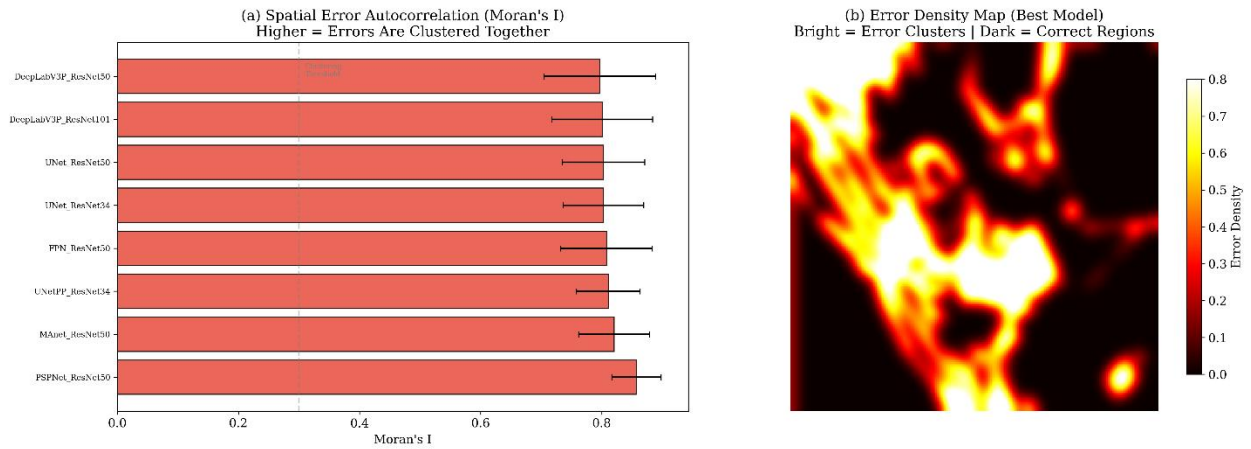


Figure 11: Autocorrelation of spatial errors. (a) Moran's I per model (bars: mean over test images; whiskers: ± 1 standard deviation; all pseudo $p < 0.001$). (b) The best model's error density map (Gaussian KDE); the bright regions denote error centres.

5.13 Class Transition Error Analysis

The most error-prone class transitions at boundaries under the ≥ 500 -pixel support threshold of Section 3.5.5 are listed in Table 11, and the full directional transition structure for the best model is depicted in Figure 12. Boundary pixels are misclassified at rates of roughly 36–97% depending on the class pair far above the interior error rates of Section 5.6 with transitions into Water the most error-prone among the tabulated pairs (Other→Water, 97.1%). Among high-support transitions, Other→Impervious (69.0% over 25,816 pixels) and Building→Impervious (74.6% over 5,131 pixels) dominate the total boundary error mass. The analysis is direction-aware and reveals systematic asymmetries: Building→Vegetation boundary pixels fail at 83.6% versus 74.8% for the reverse Vegetation→Building direction, and Building→Impervious (74.6%) far exceeds Impervious→Building (41% in the full transition matrix of Figure 12a), indicating that models preferentially expand large homogeneous classes into their smaller structured neighbours at shared edges. Such asymmetries suggest that boundary-aware losses could be weighted directionally.

Table 11. Most error-prone directional class transitions at object boundaries (best model), restricted to transitions with ≥ 500 supporting boundary pixels.

From Class	To Class	Error Rate (%)	Boundary Pixels
Other	Water	97.1	921
Building	Vegetation	83.6	1,377
Other	Vegetation	80.4	6,431
Vegetation	Building	74.8	1,379
Building	Impervious	74.6	5,131
Building	Low Vegetation	71.7	501
Other	Impervious	69.0	25,816
Other	Building	68.7	2,478
Impervious	Water	63.7	1,044

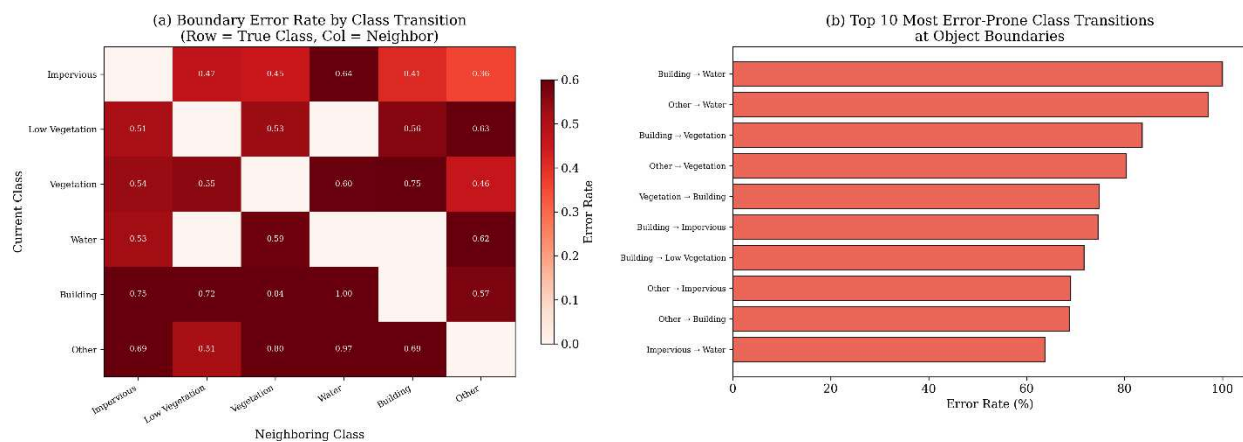


Figure 12: Class transition error analysis. (a) Directional boundary error-rate matrix (row = true class, column = neighboring class). (b) Top most error-prone transitions.

5.14 Prediction Diversity and Minimum Detectable Object Size

Pairwise disagreement analysis (Supplementary Fig. S12) shows mean disagreement of 9.9–13.3% between model pairs. PSPNet is by far the most diverse predictor (13.28% average disagreement; up to 13.8% against DeepLabV3+ ResNet-101), consistent with its architecturally distinct global-pooling design and its divergent error profile, while FPN is the most consensual (9.85%). The most similar pair is the encoder-sibling pair sharing an identical decoder DeepLabV3+ ResNet-50 vs. ResNet-101 at 8.1% indicating that decoder design, more than encoder depth, determines a model's error signature; because complementary error patterns are the raw material of ensembling (Lakshminarayanan et al., 2016), the most diverse accurate pairs identify the best two-model ensembles when the full combination of Section 5.7 is infeasible, although PSPNet's high diversity must be weighed against its low individual accuracy. Finally, the minimum detectable object size analysis (Supplementary Fig. S13) relates detection rate to ground-truth object area for three representative models: Vegetation and Other instances cross the 50% detection threshold at roughly 100–400 px, Impervious above ~500 px, and Building components remain below 50% until the largest size bins, reflecting the fragmentation of roofs into visually dissimilar sub-parts; U-Net (ResNet-50) is generally strongest on small-to-medium objects, consistent with its boundary advantage (Section 5.6). The curves for the spatially contiguous Water and Low Vegetation classes are reported for completeness only, and they are supported by a limited number of connected components. These thresholds are directly applicable to flight planning: reducing altitude results in a larger object footprint in pixels, which in turn enhances the reliability of small targets.

5.15 Summary of Key Findings

No single model dominates every dimension (Table 12): DeepLabV3+ (ResNet-101) excels in raw accuracy and interior quality; DeepLabV3+ (ResNet-50) leads in robustness and stability; FPN (ResNet-50) offers the best balanced efficiency under SEI, stably so across a broad weight range; PSPNet (ResNet-50) is the fastest and best-calibrated model, but the least accurate and among the least robust; U-Net (ResNet-50) has the best boundary quality; and U-Net (ResNet-34) is the best lightweight choice. This dispersion of strengths across architectures is the central argument for multi-dimensional evaluation: Only one of the deployment profiles below would be correct if a model were selected based solely on mIoU.

Table 12. An overview of the primary discoveries in all evaluation dimensions.

Dimension	Key Finding	Best Model
Overall accuracy	mIoU 0.7172; OA 92.33%	DeepLabV3+ (ResNet-101)
Efficiency (SEI-Balanced)	Best accuracy–speed–size balance (50.95)	FPN (ResNet-50)
Speed	622.8 FPS; 1.6 ms latency	PSPNet (ResNet-50)
Lightweight (SEI-Lightweight)	24.4M params at 383.3 FPS (35.59)	U-Net (ResNet-34)
Robustness	Lowest mean corruption drop: 20.4%	DeepLabV3+ (ResNet-50)
Calibration	Lowest ECE: 0.0397	PSPNet (ResNet-50)
Stability	Lowest pixel flip rate: 9.4%	DeepLabV3+ (ResNet-50)

Boundary accuracy	smallest gap is 41.43 pts, and the best boundary accuracy is 50.91%.	U-Net (ResNet-50)
Ensemble	relative mIoU increased by 0.7% (0.7223)	8-model probability averaging

6. Discussion

6.1 Key Findings and Implications

Conventional evaluation practices are challenged by four discoveries. Initially, the highest-mIoU model is not universally optimal. Under balanced priorities, FPN (ResNet-50) surpasses DeepLabV3+ (ResNet-101) after efficiency is accounted for through SEI. Rankings further invert in speed- and size-focused scenarios (Table 8). The weight-sensitivity analysis (Table 9) demonstrates that these reversals are stable properties of the trade-off surface, rather than artefacts of specific weights. It is argued that deployment-conditioned scores should be reported in addition to basic accuracy. Second, the 20–27% mIoU loss under realistic corruptions raises genuine reliability concerns for uncontrolled operating environments. Additionally, robustness is not a function of encoder depth or accuracy; the most robust model is DeepLabV3+ with the shallower ResNet-50 encoder. Consequently, the per-corruption vulnerability profiles of Figure 8, which are dominated by noise-family degradations, offer a concrete roadmap for robustness-targeted augmentation. Third, the statistically significant spatial clustering of errors (Moran's $I = 0.80\text{--}0.86$, $p < 0.001$) demonstrates that errors are systematic, boundary-concentrated structures rather than random noise, which opens up research directions in spatially aware losses and selectively applied refinement modules that spend computation only where errors live. Fourth, the temporal-consistency hazard for video-based drone applications that static benchmarks entirely conceal is disclosed by even sub-degree, few-pixel viewpoint perturbations, which flip 9–11% of predictions. The coincidence of these flips with boundary bands links stability and boundary quality as two faces of the same weakness.

Two additional observations warrant further attention. Confidence-consuming pipelines cannot presume that the strongest model is also the most trustworthy, as accuracy and calibration are decoupled: the least accurate model (PSPNet) is the best calibrated, and the most accurate is mid-pack. And the diversity analysis shows that decoder design, rather than encoder depth, determines a model's error signature the two DeepLabV3+ variants disagree on only 8.1% of pixels despite different encoders which explains why the eight-model ensemble, whose members share substantial decoder-level similarity, yields a real but modest +0.7% relative gain.

6.2 Practical Recommendations

The framework supports the following deployment guidance: for maximum-accuracy applications (detailed urban mapping), DeepLabV3+ (ResNet-101) is recommended where computational resources permit; for balanced deployments, FPN (ResNet-50) offers near-best accuracy at 292 FPS and modest size, and its recommendation is robust to the exact priority weighting (Table 9); for extreme-throughput or edge settings, PSPNet (ResNet-50)

provides 622.8 FPS and the best calibration, provided its lower accuracy and robustness are acceptable, with U-Net (ResNet-34) the best lightweight all-rounder; for adverse-weather operations, DeepLabV3+ (ResNet-50) is the safest default given its leading corruption robustness and prediction stability, with noise-type augmentation prioritized in training; and for safety-critical applications, probability-averaging ensembles combined with uncertainty and agreement mapping (Sections 5.5 and 5.7) should be used to flag regions requiring human verification.

6.3 Limitations

Several limitations should be acknowledged, and we state them plainly. First, each configuration was trained once under a fixed-seed protocol (Section 4.1); consequently, sub-point mIoU differences among the mid-ranked models (Section 5.1) should be read as indicative rather than significant, and we have restricted our conclusions to margins and cross-dimensional patterns large enough to be meaningful under single-run evaluation. A related observation is that validation and test rankings do not coincide perfectly U-Net (ResNet-50) attained the highest validation mIoU (0.7324) yet ranks mid-pack on the test set which is consistent with checkpoint-selection variance on a 60-image validation split and reinforces the caution above; multi-seed replication is the natural remedy and is planned as future work. Second, the evaluation uses a single dataset, and findings may not generalize to all drone scenarios (different altitudes, sensors, or geographies). Third, the corruption analysis relies on synthetic degradations that approximate but do not perfectly replicate real-world conditions, and to keep the 10×8 corrupted evaluations tractable it employs a resized-image protocol whose clean baseline differs from the tiled-patch protocol of the main results; relative degradations are therefore comparable across models, but absolute corrupted mIoU values should not be compared directly with Table 4. Fourth, several boundary-level statistics rest on modest supports even after the ≥ 500 -pixel filter, and the object-size curves for spatially contiguous classes are indicative only. Fifth, the SEI is pool-relative by construction (Section 3.4). Finally, the 30-epoch training budget, although sufficient for validation-mIoU convergence under ImageNet pretraining, leaves open whether substantially longer schedules would alter relative rankings, and transformer- and state-space-based architectures (Ma et al., 2024; Xie et al., 2021) are not yet included; extending the benchmark to such models, additional datasets, and real corruption data is the primary direction for future work.

7. Conclusion and Future Work

This paper introduced a comprehensive multi-dimensional evaluation framework for semantic segmentation of UAV aerial imagery that surpasses the conventional reliance on mIoU. The framework provided a comprehensive evaluation that encompassed the full complexity of model selection for real-world drone applications by conducting experiments on eight model configurations across fourteen evaluation dimensions. The corruption robustness benchmark quantified 20–27% degradation under adverse conditions with strong architectural variation, while the Moran's I analysis established, with statistical significance, that segmentation errors cluster

systematically at class boundaries. The proposed Segmentation Efficiency Index exposed deployment-dependent rankings that accuracy alone conceals and demonstrated their stability under weight perturbation. A consistent conclusion was reached through complementary analyses of stability, calibration, transition asymmetries, object-size detectability, uncertainty, and ensemble diversity: no singular architecture is dominant. DeepLabV3+ (ResNet-101) leads in accuracy (mIoU 0.7172; OA 92.33%), DeepLabV3+ (ResNet-50) in robustness (20.4% corruption drop) and stability (9.4% flip rate), FPN (ResNet-50) is the best all-round efficient choice under SEI, PSPNet (ResNet-50) pairs the highest speed (622.8 FPS) with the best calibration (ECE 0.0397), and an eight-model probability-averaging ensemble achieves 0.7223 mIoU, surpassing every individual architecture.

Future research will (1) extend the framework to transformer- and state-space-based architectures such as SegFormer and Mamba-family models; (2) evaluate additional UAV datasets covering diverse regions and imaging conditions; (3) perform multi-seed replication with statistical significance testing of model rankings; (4) incorporate corruption data captured under genuine adverse conditions; (5) explore architecture-specific robustness enhancement guided by the vulnerability profiles identified here; and (6) design spatially aware, directionally weighted boundary losses that exploit the clustered, asymmetric error structure revealed by the spatial autocorrelation and transition analyses.

CRedit authorship contribution statement

Nadeem Gul: Writing – review & editing, Validation, Conceptualization and Methodology

Meng Qin, Aamir Naveed Abbasi, Sichun Zheng and Gu Yuanyuan: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Changqing Cao: Supervision, Project administration

Declaration of competing interest:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Supplementary Material

This paper is accompanied by the following supplementary figures (source file in parentheses):

Fig 4. S1. Performance versus model complexity (mIoU vs. parameters).

Fig 6. S2. Macro-averaged precision, recall, and F1-score for all models.

Fig 7. S3. Heatmap of per-class F1-scores for all eight models; background column is displayed for completeness only.

Fig 8. S4. DeepLabV3+ (ResNet-101) one-vs-rest ROC curves with per-class AUC values.

Fig 10. S5. Visual error analysis: input, ground truth, prediction, and binary error map.

Fig 11. S6. Maps of pixel-level uncertainty (prediction entropy) for the optimal model.

Fig 12. S7. Ensemble prediction and multi-model agreement maps.

Fig 14. S8. Ensemble analysis: per-class comparison of probability-averaging and majority-voting ensembles versus individual models.

Fig 15. S9. Box plots depicting the class-wise IoU distributions of all eight models.

Fig 18. S10. Visual representation of the influence of corruptions on predictions and inputs.

Fig 20. S11. Resolution sensitivity: per-model sensitivity score and mIoU vs. input resolution.

Fig 24. S12. Diversity score and pairwise disagreement matrix for model prediction.

Fig 25. S13. Minimum detectable object size: detection rate versus object area per class for three representative models.

References

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2), 93–115.

<https://doi.org/10.1111/j.1538-4632.1995.tb00338.x>

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., & Shi, Z. (2024). RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–17. <https://doi.org/10.1109/TGRS.2024.3356074>

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention Mask Transformer for Universal Image Segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1280–1289. <https://doi.org/10.1109/CVPR52688.2022.00135>

Cheng, J., Deng, C., Su, Y., An, Z., & Wang, Q. (2024). Methods and datasets on semantic segmentation for Unmanned Aerial Vehicle remote sensing images: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211, 1–34. <https://doi.org/10.1016/j.isprsjprs.2024.03.012>

Fan, T., Wang, G., Li, Y., & Wang, H. (2020). MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation. *IEEE Access*, 8, 179656–179665. <https://doi.org/10.1109/ACCESS.2020.3025372>

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On Calibration of Modern Neural Networks* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1706.04599>

- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
<https://doi.org/10.1109/CVPR.2016.90>
- Hendrycks, D., & Dietterich, T. (2019). *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1903.12261>
- Iakubovskii, I., P. (2019). Segmentation Models Pytorch [Computer software]. *GitHub*.
https://github.com/qubvel/segmentation_models.pytorch
- Iakubovskii, P. (2019). *Segmentation Models Pytorch* (Version 0.3.4) [Computer software]. *GitHub*.
https://github.com/qubvel/segmentation_models.pytorch
- Institute of Computer Graphics and Vision (ICG), Graz University of Technology* (ICG, Graz University of Technology, 2019). (2019). [Dataset].
- Kamann, C., & Rother, C. (2020). Benchmarking the Robustness of Semantic Segmentation Models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8825–8835.
<https://doi.org/10.1109/CVPR42600.2020.00885>
- Kendall, A., & Gal, Y. (2017). *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1703.04977>
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003. <https://doi.org/10.1109/ICCV51070.2023.00371>
- Krähenbühl, P., & Koltun, V. (2012). *Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials*. <https://doi.org/10.48550/ARXIV.1210.5644>
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2016). *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.1612.01474>
- Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. <https://doi.org/10.1109/CVPR.2017.106>

- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. <https://doi.org/10.1109/ICCV48922.2021.00986>
- Loshchilov, I., & Hutter, F. (2017). *Decoupled Weight Decay Regularization* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.1711.05101>
- Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., & Yang, M. Y. (2020). UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165, 108–119. <https://doi.org/10.1016/j.isprsjprs.2020.05.009>
- Ma, X., Zhang, X., & Pun, M.-O. (2024). RS³ Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5. <https://doi.org/10.1109/LGRS.2024.3414293>
- Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., & Brendel, W. (2019). *Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1907.07484>
- Moran, P. A. P. (1950). NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. *Biometrika*, 37(1–2), 17–23. <https://doi.org/10.1093/biomet/37.1-2.17>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1912.01703>
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). *SAM 2: Segment Anything in Images and Videos* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2408.00714>

- Safavi, F., & Rahnemoonfar, M. (2023). Comparative Study of Real-Time Semantic Segmentation Networks in Aerial Images During Flooding Events. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 4–20. <https://doi.org/10.1109/JSTARS.2022.3219724>
- Wang, J., Zheng, Z., Ma, A., Lu, X., & Zhong, Y. (2021). LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation (Version 6). arXiv. <https://doi.org/10.48550/ARXIV.2110.08733>
- Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 196–214. <https://doi.org/10.1016/j.isprsjprs.2022.06.008>
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2105.15203>
- Xu, J., Xiong, Z., & Bhattacharyya, S. P. (2023). PIDNet: A Real-time Semantic Segmentation Network Inspired by PID Controllers. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19529–19539. <https://doi.org/10.1109/CVPR52729.2023.01871>
- Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11217, pp. 334–349). Springer International Publishing. https://doi.org/10.1007/978-3-030-01261-8_20
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230–6239. <https://doi.org/10.1109/CVPR.2017.660>
- Zhu, Q., Cai, Y., Fang, Y., Yang, Y., Chen, C., Fan, L., & Nguyen, A. (2024). Samba: Semantic segmentation of remotely sensed images with state space model. *Heliyon*, 10(19), e38495. <https://doi.org/10.1016/j.heliyon.2024.e38495>

